United Nations ECE/CES/2015/30

Economic and Social Council

Distr.: General
9 April 2015

Original: English

**Economic Commission for Europe**

Conference of European Statisticians

**Sixty-third plenary session**
Geneva, 15-17 June 2015
Item 4 of the provisional agenda
**Modernisation of statistical production and services and managing for efficiency**

# Statistical survey management platform

**Note by the Turkish Statistical Institute**[1]

*Summary*

This paper presents a recently developed statistical survey management platform designed by the Turkish Statistical Institute to standardize statistical production, increase the quality of statistics and decrease the workload. The platform generates generic applications for operational phases of surveys, in line with the Generic Statistical Business Process Model. Currently, 83 of the total of 92 surveys carried out by Turkish Statistical Institute have been moved onto this platform.

The paper is presented for discussion to the second session of the Conference of European Statisticians' seminar "Modernisation of statistical production and services and managing for efficiency".

[1] This document was submitted late due to delayed inputs from other sources.

GE.15-07350 (E)

*1507350*

Please recycle

# I. Introduction

1. The Turkish Statistical Institute (TurkStat) is the producer and coordinator of the official statistical system in Turkey with a mission to produce high quality statistics. TurkStat recently developed a statistical survey management platform, called Harzemli, to standardize statistical production, increase the quality of statistics and decrease the workload.

2. The Harzemli Platform is based on the Data Documentation Initiative (DDI) and provides an environment for developing generic applications. The system uses a DDI file containing reference and structural metadata information of the survey as an input, and then it generates the survey data entry application with a generic user interface. This results in a desired state where the web, desktop and mobile interfaces for different surveys have the same look and feel as well as the same underlying infrastructure.

3. The Harzemli Platform is an innovative statistical production platform that is currently used in the production of official statistics in Turkey. Before the Harzemli Platform, staff at the Information and Communication Technology (ICT) Department had been assigned the implementation of separate web applications for each survey. Previously, different statistical surveys relied on various software products with different logic, the applications looked and felt different and there was no standard coding approach. The Harzemli Platform has provided a high level of standardization on data collection applications, thus, generating low-cost, better-quality and long lasting applications for all phases of statistical production from data collection to data dissemination.

4. The Harzemli Platform is a tool for the standardization and modernization of the statistical production process. It increases the reliability and consistency of official statistics. The platform also minimizes the time used and provides standards to allow effective production and dissemination of data. It standardizes the names of all variables which helps improve coherence, enables software development with standard application codes, and is not dependent on individual IT persons. Furthermore, a significant outcome introduced by the new platform is the shorter and faster software development process.

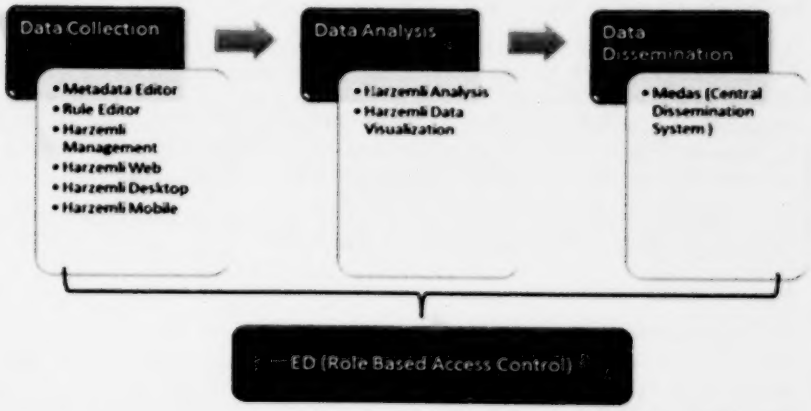
5. The new platform has also improved the dissemination phase, as defined in the Generic Statistical Business Process Model (GSBPM). The end users of statistics are able to make dynamic queries on TurkStat's website in addition to using static tables and statistical releases. Turkstat's new dissemination system, called MEDAS, consists of a common database and a common application for using it. The system allows querying all different statistical subjects at the same time. This unique feature can lead the users to find new combinations of statistical issues.

# II. Description of current work

## A. System components

6. The Harzemli Platform consists of a Role-Based Access Control (RBAC) implementing Identity Management (IDM), Rule Editor, Desktop, Web, Mobile, Management Console, Analysis, Data Visualization and Dissemination projects.

Figure 1
**Harzemli Platform System Components**




### 1. Metadata Editor

7. The Metadata Editor is software used to create the metadata according to international metadata standards (DDI and Dublin Core). It enables entering both the reference and the structural metadata and also allows storing entered metadata as Extensible Markup Language (XML) files in DDI format.

### 2. Rule Editor

8. The Rule Editor is a desktop application used in defining the rules for a field or rules in defining the interactions between the fields in the electronic data collection forms. It can be used to specify simple edit rules, *show* or *hide* fields based on conditions, to make the fields read only, to specify function rules and nullable field rules.

9. The output XML file of the Rule Editor is used as an input to Harzemli Web, Harzemli Desktop, Harzemli Mobile and Harzemli Analysis applications. Using this output XML file throughout data collection phase ensures that erroneous data are not allowed inside input database and controls the flow of the survey.

### 3. Harzemli application platforms

10. Harzemli consists of 3 application platforms as follows:

*(a) Harzemli Desktop*

11. The Harzemli Desktop is a desktop application that is designed to work offline (without any need for internet) to collect data using netbooks/notebooks. The Desktop application is the first data collection project designed in Harzemli among the three data collection tools (Desktop, Web and Mobile) available in the Harzemli platform. It is designed to move from paper based data collection to electronic data collection. It receives the DDI and rule XML files as input and generates the data collection user interfaces for surveys. The main benefit of Harzemli Desktop is that it allows the field interviewers to collect data electronically without an internet connection. When the internet connection becomes available, application synchronizes the off-line content with the central database in a bidirectional manner.

*(b) Harzemli Web*

12. Harzemli Web is a web application that transforms survey questionnaires prepared in DDI format into the respective data collection applications.

13. With guidance from the Metadata and Standards Department and the Information Technologies Department, the subject matter departments design the layout and appearance of their survey forms. Harzemli Web dynamically generates the survey forms through processing DDI and Rule XML files. Production units decide the design of the web pages by preparing these files. Units decide on how many pages the survey should be, how many sections it will have, the length of the textboxes, the usage of buttons and so on.

14. Harzemli Web has two perspectives, standard perspective and analysis perspective, which provide data collection and data analysis functionalities.

*(c) Harzemli Mobile*

15. Harzemli Mobile is an Android application that is designed to be used in tablets. The aim of Harzemli Mobile is to take advantage of mobile operating systems and light weight devices in the field. Similar to Harzemli Desktop, by using DDI and rule XML files as input, Harzemli Mobile generates data collection applications dynamically. Harzemli Mobile is able to save data offline without any need for internet connection. When the internet connection becomes available it is possible to transfer the data from tablets to central databases through web services.

## 4. Harzemli Management Console

16. Harzemli Management Console is a web application that is used for authentication, authorization, notification through text messages, and for reporting and planning surveys. Harzemli Management Console automatically generates the necessary database tables using the DDI and rule XML files for surveys as input. It is also used for managing the sample. Any special need with respect to surveys is covered by Harzemli Management Console.

## 5. Role-Based Access Control

17. Role-Based Access Control is an application that provides access to programs/applications with hierarchical role-based authorization.

18. The types of roles in this system are:

- The main roles of the unit: It is given by default to the unit manager.
- Sub-roles: Roles that are created by the unit manager.

19. Unit manager can authorize the staff by creating several sub-roles. The system is integrated with human resources system. The authorizations are given to the management roles in the hierarchy. Field survey staff and external survey respondents can be authorized by the managers using the system.

## 6. Harzemli Analysis

20. Harzemli Analysis is a data analysis tool that allows the users to perform different analysis on data through different technologies. Users of this analysis tool are able to create and run their own error finder rules through automatic transformation from logical expressions into the Structured Query Language (SQL) which is a standard language for accessing databases. The user may also run special queries (through automatic transformation from XML based rules to SQL statements) and analysis on files from other statistical analysis systems (including the Statistical Package for the Social Sciences

(SPSS), the Statistical Analysis Software (SAS) and the R-programming language) for advanced analytics.

### 7. Harzemli Data Visualization

21. Harzemli Data Visualization tool allows statistical analysis with the graphics created by using the data collected by the Harzemli Platform. With this application, SQL sentence which is generated according to the table selected by the user and column values of this table are sent to the (R) server. Visualization of data is implemented according to the data type and parameters selected by user.

### 8. Data dissemination (MEDAS)

22. The number of printed statistical publications is reducing. The fact that the Statistical Yearbook, prepared by Statistics Canada, will no longer be published is an example of this development. Electronic dissemination may be the main publication channel in the near future. Typically statistical institutes use various electronic dissemination channels: press releases, predefined tables, dissemination databases (databanks), microdata files, thematic publications, etc. The dissemination databases on the TurkStat website can be defined as an aggregated data repository that is populated with data from microdata warehouses. Microdata warehouses are often considered as service developed for researchers, academics, data miners or for other heavy users of statistics. Known advantages of databases can be classified as; avoiding duplication of information, helping to develop web applications, security, quick access to updated data, multi-language possibilities, and enforced provision of metadata.

23. TurkStat has had a dissemination database for more than a decade. However, the strategy for adding new statistics in the dissemination database has been to find a tailored solution for each case with the ICT Department which assigns a database specialist to carry out the task. The specialist would hold some meetings with the statistical subject matter unit in question, and then prepare tables specifically designed for the statistic in question. The new table would not relate to other statistics directly and there would not be any central metadata repository that could lead the end user to the related statistical topics. Hence, the end user must work hard to be able to combine different statistics together for analysis, and will have to open different tables and reports and merge the data manually into a single spreadsheet.

24. One other handicap of the old system was the web application part. As mentioned above, each statistical topic has its own database tables applying different structures. This not only brought maintenance problems to the ICT Department, but also resulted in developing different Java applications for each statistic, which added maintenance problems and is highly dependent on individual IT specialists. End users had to deal with various web applications although the ICT Department worked hard to provide an interface that looks as if there was just one single application behind. However, there was a lack of consistency between the applications that was also visible when querying in the different databases.

25. Moreover, making new statistics available to users involved a manual process. The subject matter unit staff contacted the database specialist saying that the related press release is released and data can be shared with end users. After that, the database specialist loaded the final data to the dissemination database. This required various different data flows, wikis and codes for each statistic. In addition, one of the main reasons for the change was that the old technology used for dissemination databases was vulnerable to security threats.

26. The ICT Department conducted a study on possibilities to standardize the above-mentioned processes without changing the content. The database design was the challenging part, since all queries into the databases required first selecting a specific statistical topic. Instead, the new approach should be generic and allow the user to see a single report regardless of the number of statistical topics chosen. The data should be provided in modern pivot tables.

27. The role of metadata for internal processes is considered important in TurkStat, especially since 2012. TurkStat has been using a classification server for a decade. After analysing its data structure, TurkStat decided that the metadata of MEDAS, the new data dissemination system, could be based on the existing classification server. Some statistical topics were chosen as pilot areas for the new system. The results of these pilot projects were considered quite satisfactory. MEDAS is now in use since April 2014. Hence, TurkStat is now migrating its old-fashioned dissemination databases to MEDAS. All new statistics that will be provided via the databank are developed using MEDAS. Workshops have been held to inform the staff of subject matter units about the new data structure of MEDAS.

28. To sum up the key impact of the new databank system with MEDAS:

(a) The ICT Department builds the framework and the data are filled in by the subject matter units. This reduces the burden on ICT staff and ICT directors. There is only one dissemination schema and one database design that includes all the dissemination data. There is only one databank application. However, some other practical applications are and will be written based on MEDAS data. MEDAS reports reduce the reporting burden on ICT staff;

(b) Manual interventions in the data dissemination phase have been reduced, and there is no longer a need to wait for a call or email from the subject matter unit saying that the data could be uploaded to the web. MEDAS waits for the related press release (if any) to be released and the data are shown to the end users after the release data;

(c) Users are now able to use modern pivot reports and compare any number of subjects in the same report page;

(d) The uniform data model makes it easy to develop web services. MEDAS, especially with its data design, is an innovation for statistical production processes and some developments are being conducted for serving the data with graphs and thematic maps.

## B. Objectives

29. The main objectives of the Harzemli Platform are

- to shorten the process of coding for data entry programme;
- to reduce dependency on developers when developing data entry programmes;
- to develop data entry programme with standard application codes;
- to standardize the names and code lists of all the variables for integrity of data;
- to ensure faster compilation of private sector data entry by respondents in the survey and the immediate correction of erroneous input data.

30. The model could be a reference to other public institutions for their statistical production processes.

## C. Implementation steps

### 1. Feasibility study

31. After analyzing an XML based generic application development platform, named "wizard", TurkStat decided to launch the Harzemli Project, an advanced version of this "wizard" application.

### 2. Cost-benefit analysis

32. A cost- benefit analysis was conducted on the potential benefits and outcomes of the project. It was decided that instead of buying software or outsourcing, it would be more feasible to create our own software with the efforts of staff at the ICT Department.

### 3. Risk management plan

33. Risk prevention plan has been developed based on risk assumptions. The underlying conditions are determined using expert knowledge and experiences gained during the development of wizard application.

### 4. Quality management plan

34. Quality management plan has been defined with the aim of using DDI xml format in an efficient way and benefitting from recent development of the DDI standard.

### 5. Change management plan

35. Change management plan has been developed based on job steps which are required to be made and followed when a revision request made about survey studies.

### 6. Human resources management plan

36. Human resources management plan consists of defining project roles, assigning responsibilities and designing project organization schema. The number of project team members was increased step by step by using the human resource management plan after implementing the first few modules of the project successfully.

### 7. Disclosure and analysis of requirements

37. At the beginning of the project, the labour force survey was selected as a pilot study. In order to analyze processes and to understand application requirements in detail, a joint working group of software developers, subject matter unit staff and users of the application was set up.

38. System and user requirements (functional, non-functional and quality requirements) are defined based on the meeting decisions regarding business logic and data structure of the application. These meetings are organized a way that requests from the end users can also be incorporated into the general requirements.

### 8. Project management

39. After completion of planning and requirement analysis, tasks were assigned and development phase of the project started. In the management of the project, agile project management methods are used so that the project modules can be extended based on the user requirements. First, the Harzemli Desktop module was developed. Based on the experiences gained in the development of the Harzemli Desktop, Harzemli Web and Harzemli Mobile modules were designed. The most complex and error prone requirements and components were identified to be developed first to detect the possible errors at the

beginning of the project. The pilot, the labor force survey, was selected because it contains many of the requirements that the other surveys have and has a complex workflow.

### 9. Defining changing requests and implementation of approved ones

40. In the quality control testing phase, tests were carried out to control whether the requirements are met properly. Any errors were detected and flagged to be fixed. In order to mitigate the risks, in order to prevent collecting incorrect or missing data, paper forms were continued to be used in parallel in the first few pilot studies. The most complex studies are implemented by using the Harzemli Project in developing Harzemli modules.

### 10. System test

41. The ICT Department manages the testing activities by defining test scenarios and tracking these test cases. Survey on Information and Communications Technologies (ICT) Usage in Households was the first study implemented fully by Harzemli Desktop. It was tested by staff responsible with defining survey requirements in the statistical unit, software test professionals in the ICT Department and by staff responsible for collecting data via the application. After completion of the testing phase, most of the bugs and inaccurate workflow of the application were fixed by the project team before the application was ready to be used in real environment. Each new release of application is tested by both testing staff and statistical unit staff. In order to run a performance test on the application which is used by lots of users, it is tested by simulating users generated by test simulation tools.

## D. Project milestones and results

42. The project milestones are as presented below:

Table 1
**Harzemli Platform Project Milestones**

| | Milestones of the project | Development year |
|---|---|---|
| 1 | Harzemli Desktop | 2012 |
| 2 | Harzemli Rule Editor | 2012 |
| 3 | Harzemli Management Console | 2012 |
| 4 | IDM | 2013 |
| 5 | Harzemli Web | 2013 |
| 6 | MEDAS | 2014 |
| 7 | Harzemli Mobile | 2014 |
| 8 | Harzemli Analysis | 2014 |

**Results**

43. In 2013, 32 surveys were implemented in the Harzemli Project. Out of these 32 surveys, 6 were designed in Harzemli Desktop and 26 were designed in Harzemli Web. As of 2014, the total number of surveys moved to Harzemli is 82. In 2015, 22 more surveys will be moved to Harzemli (see table 2).

Table 2
**Results of Harzemli Project**

| | Harzemli Desktop | Harzemli Web | Harzemli Mobile | |
|---|---|---|---|---|
| 2013 | 6 | 26 | * | 32 |
| 2014 | 4 | 46 | * | 50 |
| 2015 | 2 | 13 | 7 | 22 |
| TOTAL | 12 | 85 | 7 | 104 |

# III. Conclusion

44. As the Harzemli Platform uses common classification and code lists, it not only standardizes data collection and dissemination processes, but also shortens the duration of preparation time for data entry programs from 40 days to 10 days, and shortens the time needed for preparing statistical releases for a minimum of 4 days.

45. So far, the Harzemli Platform has proven to be a great success and has been strongly supported by both the management and staff. The Harzemli Platform will be enlarged in the near future by adding new modules that are currently being developed.

## A. Productivity gains

### 1. Time

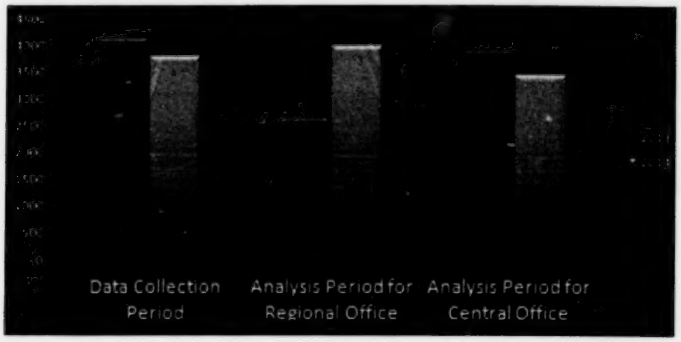
*(a) Decreased time for software engineers to develop data entry applications*

46. For every survey conducted by TurkStat, study-specific software development has been eliminated. Depending on the type of the survey, data entry programmes are now automatically created through one of the three data entry applications (Harzemli Desktop, Web or Mobile applications) based on XML files used for defining the metadata and rules (i.e. DDI.xml and Rule.XML). While the average time necessary for developing a data entry programme for a statistical survey with an average level of difficulty was about 4 weeks, it is now only a week.

*(b) Decreased time for data collection*

47. Before the Harzemli Platform, surveys were printed and sent to the respondents to collect data. Data entry was performed in regional offices through processing forms that had previously been filled in the paper by respondents. Using the Harzemli Platform data are directly collected from the respondents, thus saving time and labour. Data collection time is shortened by 8 per cent and because of instant access to data, analysis time available for regional offices has been increased by 50 per cent. As data come directly from respondents to TurkStat and from the regions, regional offices are now able to spend more time on data processing and analysis. This has led to a 12 per cent decrease in the analysis time that is necessary for the officers at the central office.

Figure 2
**Time for data collection and analysis**




*(c) Decreased time for preparation of press releases*

48. Having the collected data transmitted instantly to the central office and diminishing the necessary time for analysis, data dissemination has been quickened. The time period necessary to prepare the press releases has been shortened. For example, the time period that is necessary for the preparation of the monthly labor statistics statistical releases has been shortened by 4 days.

## 2. Quality

*(a) Software process standardization*

49. With the Harzemli Platform, standard business processes have been established and job descriptions have been updated. Thus, control, consistency and order of processes are established and complexity is reduced. Dependency of the software process on individual persons has been reduced.

*(b) Data integrity*

50. In order to generate data with the Harzemli Platform, usage of standard code lists and standard variable definitions ensure better coherence of data. Code lists and variables were not common for usage for each study before the Harzemli project. Through standardization, possibilities for data matching between surveys have improved. Thus, it also made possible the development of data exchange and distribution software at TurkStat.

*(c) Common code development*

51. Before the Harzemli Platform, applications for each survey were developed using different programming languages and technologies. Therefore, it was almost impossible to inspect the quality of all the applications. Prior to the Harzemli project, maintenance, quality assurance and software testing were done for 90 separate data entry applications whereas now this has been reduced to only 3 applications (namely, Harzemli Desktop, Harzemli Web and Harzemli Mobile). Switching from individual software development to unified development process, more functional, reusable, multi-purpose applications with standardized interface have been developed.

*(d) Common components available*

52. Software code written prior to the Harzemli project has been split into modules as the project has progressed and layering has been implemented. The software components

have been optimized to be suitable for being commonly used in Harzemli Desktop, Web and Mobile applications and, thus, the costs of separately writing and maintaining components have disappeared. Having metadata standardization done and the standard reporting mechanism accomplished under, the need to write separate programmes for each statistics is eliminated.

### 3. Costs

*Paper*

53. In order to keep pace with constantly evolving technology, and to further reduce manual labour and costs to produce data in a timely and efficient manner, TurkStat is already carrying out all business surveys through the Harzemli Web application. Paper editions of these surveys have been cancelled. In 2014, thanks to the Harzemli Project, 7 565 680 sheets of paper were saved for all surveys targeted at a total of 756 568 respondents.

54. In 2015, the sample volume has reached the level of 2 045 974 respondents through Harzemli Web and Harzemli Mobile applications. Considering that an average of 10 pages are needed for conducting a survey with a respondent, 20 459 740 sheets of paper savings are anticipated.